Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2017)

Выпуск 16

Том 1 из 2
Компьютерная лингвистика:
практические приложения

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2017)

Issue 16

Volume 1 of 2
Computational Linguistics: Practical Applications

УДК 80/81; 004
ББК 81.1
К63

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (Москва, 31 мая — 3 июня 2017 г.). Вып. 16 (23): В 2 т. Т. 1 — М.: Изд-во РГГУ, 2017.

Сборник включает 71 доклад международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2017», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

Предисловие

16-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 23-й международной конференции «Диалог». На основании мнений наших рецензентов для публикации в ежегоднике Редсоветом был отобран 71 доклад из числа примерно ста работ, которые были рекомендованы по результатам рецензирования для представления на конференции в 2017 году.

Работы в сборнике отражают все основные направления исследований в области компьютерного моделирования и анализа естественного языка, представленные на конференции:

- Компьютерные лингвистические ресурсы
- Компьютерный анализ документов (классификация, поиск, анализ тональности и т.д.)
- Корпусная лингвистика (создание, разметка, методики применения и оценка корпусов)
- Лингвистические онтологии и автоматическое извлечение знаний
- Лингвистический анализ Social media
- Лингвистический анализ речи
- Машинный перевод текста и речи
- Модели и методы семантического анализа текста
- Модели общения
- Теоретическая и компьютерная лексикография
- Типология и компьютерная лингвистика
- Формальные модели языка и их применение в компьютерной лингвистике

В соответствии с традициями «Диалога», старейшей и крупнейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом и моделированием. Одной из важнейших целей конференции была и остается поддержка создания современных компьютерных ресурсов, моделей и технологий для русского языка.

В годовом цикле проведения конференции в рамках программы Dialogue Evaluation проводится тестирования технологий решения отдельных задач компьютерного анализа языка. На конференции подводятся итоги проведенных тестов, а статьи организаторов и наиболее успешных участников представляются в настоящем сборнике.

В этом году было проведено два тестирования:

1. По идентификации внешних заимствований (External Plagiarism Detection)
2. По оценке методов морфологического анализа русского языка, с акцентом на тексты Social Media.

Как обычно, результатом проведенных тестирований стали не только объективные данные о качестве работы различных методов и алгоритмов, но также и открытые для использования эталонные размеченные корпуса, т. н. золотые стандарты, позволяющие любым исследователям проводить сравнительные оценки эффективности своих технологий.

Все направления «Диалога» важны, но каждый год какие-то темы занимают особое место в программе конференции и в составе ежегодника. В этом году можно назвать две таких темы:

1. Применение методов глубинного машинного обучения: прежде всего — нейросетей и таких результатов их применения как word embeddings, как для прикладных задач, так и в лингвистических исследованиях.
2. В программе конференции этого года особенно заметны работы по использованию параллельных корпусов для лингвистических исследований. Такие корпуса уже давно и успешно используются в NLP, например, для обучения статистических моделей машинного перевода, автоматической дизамбигуации, автоматического построения языковых моделей. Но параллельные корпуса оказываются также и важным инструментом контрастивных лингвистических исследований.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.

- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что его бумажный вариант, который вы держите в руках, является вторичным по отношению к сборнику, который размещается на сайте конференции и индексируется Scopus. Мы рекомендуем при цитировании использовать именно сетевую версию.

Программный комитет конференции «Диалог»

Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция «Диалог» проводится под патронажем Российского фонда фундаментальных исследований при организационной поддержке компании ABBYY.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Богуславский Игорь Михайлович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Буате Кристиан	Университет Джозефа Фурье — Гренобль 1, Франция
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича, Россия
Кобозева Ирина Михайловна	Московский государственный университет им. М. В. Ломоносова, Россия
Козеренко Елена Борисовна	Институт проблем информатики РАН, Россия
Корбетт Гревил	Университет Суррея, Великобритания
Кронгауз Максим Анисимович	НИУ «Высшая школа экономики», Россия
Лукашевич Наталья Валентиновна	НИВЦ МГУ им. М. В. Ломоносова, Россия
Маккарти Диана	Кембриджский университет, Великобритания
Мельчук Игорь Александрович	Монреальский университет, Канада
Нивре Йоаким	Уппсальский университет, Швеция
Ниренбург Сергей	Университет Мэриленда, Балтимор, США
Осипов Геннадий Семёнович	Институт системного анализа РАН, Россия
Раскин Виктор	Университет Пердью, США
Селегей Владимир Павлович	Компания ABBYY, Россия
Хови Эдуард	Университет Карнеги — Меллон, США
Шаров Сергей Александрович	Университет Лидса, Великобритания

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Байтин Алексей Владимирович	Компания Yandex
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ им. М. В. Ломоносова
Захаров Леонид Михайлович	Московский государственный университет им. М. В. Ломоносова
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН им. А. А. Харкевича
Кобозева Ирина Михайловна	Московский государственный университет им. М. В. Ломоносова
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	Компания Yandex
Ляшевская Ольга Николаевна	Институт русского языка им. В. В. Виноградова РАН
Толдова Светлана Юрьевна	НИУ «Высшая школа экономики»
Федорова Ольга Викторовна	Московский государственный университет им. М.В. Ломоносова
Шаров Сергей Александрович	Университет Лидса

Секретариат

Атясова Анастасия Леонидовна, *координатор оргкомитета*	Компания ABBYY
Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Гусева Анна Александровна, *координатор Dialogue Evaluation*	Компания ABBYY
Севергина Екатерина Александровна, *администратор оргкомитета*	Компания ABBYY

Рецензенты

Августинова Таня
Антонова Александра Александровна
Азарова Ирина Владимировна
Андрианов Андрей Иванович
Апресян Валентина Юрьевна
Архангельский Тимофей Александрович
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Бенко Владимир
Бердичевский Александр Сергеевич
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бочаров Виктор Владиславович
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Галинская Ирина Евгеньевна
Галицкий Борис Александрович
Гельбух Александр Феликсович
Гецевич Юрий Станиславович
Гращенков Павел Валерьевич
Губин Максим Вадимович
Даниэль Михаил Александрович
Диконов Вячеслав Григорьевич
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрушина Нина Роландовна
Зализняк Анна Андреевна
Захаров Виктор Павлович
Захаров Леонид Михайлович
Ильвовский Дмитрий Алексеевич
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Катинская Анисья Юрьевна
Клышинский Эдуард Станиславович
Кибрик Андрей Александрович
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кронгауз Максим Анисимович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лопухин Константин Александрович
Лукашевич Наталья Валентиновна
Лютикова Екатерина Анатольевна
Мисюрев Алексей Владимирович
Наков Преслав
Недолужко Анна Юрьевна
Падучева Елена Викторовна
Пазельская Анна Германовна
Паперно Денис Аронович
Панченко Александр Иванович
Переверзева Светлана Игоревна
Петрова Мария Андреевна
Пивоварова Лидия Михайловна
Пиперски Александр Чедович
Подлесская Вера Исааковна
Рахилина Екатерина Владимировна
Скулачева Татьяна Владимировна
Смирнов Иван Валентинович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
Соколова Елена Григорьевна
Сомин Антон Александрович
Сорокин Алексей Андреевич
Сорокин Виктор Николаевич
Старостин Анатолий Сергеевич
Степанова Мария Евгеньевна
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Турдаков Денис Юрьевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Шелманов Артём Олегович
Янко Татьяна Евгеньевна

Contents[1]

Приглашенные доклады

Компьютерная лингвистика: практические приложения

[*] Доклады упорядочены по фамилии первого автора в соответствии с английским алфавитом. The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

EXTRACTING CHARACTER NETWORKS TO EXPLORE LITERARY PLOT DYNAMICS

Skorinkin D. A. (dskorinkin@hse.ru)

Higher School of Economics, Moscow, Russia

In this paper we apply network analysis to the study of literature. At the first stage of our investigation we automatically extract networks (graphs) of characters for each part of Leo Tolstoy's novel *War and peace* using two different techniques for network creation. Then we evaluate these two techniques against a set of manually created gold standard networks. Finally, we use the method that demonstrated better performance in our evaluation to test a literary hypothesis about Tolstoy's novel. The hypotheses we intended to prove was that the parts of the novel describing war (i.e. those where the battlefield or military units are the primary settings), have statistically lower density of interaction between characters, resulting in lower network density, higher network diameters and lesser average node degrees. By showing this correlation we mean to demonstrate the applicability of network analysis to computational research of fictional narrative (e.g. detection of tension changes in the plot).

Key words: networks, network theory, social network analysis, literary network analysis, graph models, digital literary studies, Russian literature

1. Introduction

Over the last decades network analysis found successful applications to a great variety of fields ranging from sociology and political science to criminology and epidemiology. In recent years literary scholars, whose objects of study are also convertible to vertices and edges, turned their attention to graph[1] theory and started actively borrowing methods from social network analysis.

It has been shown that networks of fictional characters are similar to those of real social networks [Alberich et al., 2002] and share certain characteristics (e.g. power law distributions) with all other complex network types [Park, Kim, 2013]. Network theory allowed researchers to make novel observations about the composition and plot of literary pieces [Elson et al., 2010], [Moretti, 2011] and get new "insight into the roles of characters in the story" [Agarwal, Corvalan et al., 2012].

However, this ability to look at certain work of fiction from a different angle is not the only advantage of such graph-based formalization. Combined with various NLP-related techniques for automatic network extraction (some of which are implemented in this study), network analysis also opens the doors to large-scale analysis of fiction.

---

[1] In this paper we treat 'network' and 'graph' as synonymous words both meaning 'a set of vertices connected by edges'.

Such analysis, often referred to as 'distant reading' [Moretti, 2013], 'scalable reading' [Weitin, 2017] or 'macroanalysis' [Jockers, 2013], has been a point of heated debates in literary studies in recent years. The proponents of large-scale computational analysis of literature claim that close reading and precise analysis of particular pre-selected texts, traditional for literary scholars of the past, can no longer be considered sufficient for scientific research, as these approaches are only applicable to very narrow selections of works (usually the so-called *canon*, itself a very ill-defined and arguable concept). They suggest literary scholars should 'learn how *not* to read' the texts they study [Moretti, 2013] and 'start counting, graphing, and mapping them instead' [Moretti, 2007]. And although there is a fair share of criticism towards this approach, the fact remains that even a single literary movement in a single national literature usually generates more text than a single person can read, much less analyze, in his lifetime[2].

2. Related work

There has been a number of research on extraction and exploration of fictional networks. [Agarwal, Kotalwar et al., 2013] extract social events, i.e. interactions between characters or perceptions of one character by another, form Carrol's Alice in Wonderland; [Ardanuy, Sporleder, 2015] use networks to perform genre classification of XIX century novels; [Lee, Yeung, 2012] investigate the structure of the Old Testament linking people to places thus creating spatio-personal networks; [Elson et al., 2010] explore 60 British XIX century novels through conversational networks generated from dialogues of the characters. That latter work, presented at the ACL 2010 conference, deserves a separate mention. Unlike many others, [Elson et al., 2010] do not limit themselves to network extraction and evaluation against some gold standard; their main goal is to use structural properties of networks to disprove an influential literary theory(hypothesis). The hypothesis claimed that 'rural' novels reflected typical social structure of a village with its close-knit community of people familiar to each other, whereas 'urban' novels demonstrated more complex social networks with several communities, lesser overall density and a plethora of 'weak ties'; and that therefore the importance and amount of dialogue decreased as novels shifted from rural to urban settings after the industrial revolution. However, [Elson et al., 2010] did not find this to be the case.

In our investigation we also try to employ network parameters and network statistics as a means of testing a literary hypothesis. An additional motivation for our study was lack of literary network research made on Russian material, the only notable exception being [Bocharov, Bodrova, 2014]. That latter work, however, does not go beyond basic network extraction and evaluation, and its authors made no attempt to prove any literary theory or hypothesis.

---

[2] For instance, it is estimated that Victorian novels alone make up a corpus of about 60,000 texts [Moretti, 2013]

3. Hypothesis and relevant network metrics

Much like [Elson et al., 2010], we chose to study the relation between the settings in which the plot unfolds and the structural properties of the character network. However, in our case the main opposition was not 'urban' vs 'rural', but 'war' vs 'peace'. This antithesis not only gave the novel its ever-famous title[3], it is certainly among the pillars of the whole work. One of the most acclaimed Tolstoy scholars Boris Eikhenbaum spoke of *War and peace* as a novel where "The Iliad" (i.e. war) must "follow the Odissey" (i.e. peace) [Eikhenbaum, 2009 (1931), p. 497]; notable American slavist Gary Saul Morson calls this the "central opposition" of the book and claims that "the salon and the battlefield represent the extremes of order and chaos — of 'peace' and 'war' — in War and peace" [Morson, 1987, p. 97]. Note that Morson uses spatial settings — salon and battlefield — as metonymic labels for the complex concepts of 'war' and 'peace'; this indicates that spatial dynamics of narrative is the primary marker for switches between these two 'extremes'. And indeed, chapters (главы) and even entire parts (части) of Tolstoy's *War and peace* can be fairly easily subdivided into 'peaceful' and 'wartime' ones by simply looking at the space in which the plot unfolds.

This contrast between war and peace can be observed on many levels, among which the level (and intensity) of character interactions. It were changes at this level that we hoped to detect with network analysis. We had two reasons to believe that such interactions should be visibly influenced by settings:

1. Research on dramatic texts shows that tragedies tend to have lower density of networks [Trilcke et al., 2015b], and a possible explanation for this is that tragic events need less verbal interaction and verbal space than, for instance, comic scenes; this could also be the case for 'war' and 'peace' split;
2. Tolstoy's 'war narrative' is very individualistic [Morson, 1987, p. 99], it is largely focused on the inner state of a single person on the battlefield (e.g. Andrey in the Austerlitz battle; Nikolay during the Battle of Schöngrabern and the affair at Ostróvna, Pierre in Borodino).

Therefore our hypothesis was that there should be a strong correlation between the type of settings and certain standard network metrics which reflect the intensity of interactions. The metrics we propose are:

1. network density, which is the ratio of the number of edges in a graph to the maximum possible number of edges in that graph (i.e. if each node was connected to every other node).
2. network diameter, which is the length of the longest path between one node and another in that network, measured as the number of edges. Can only be calculated if there is one single component in the graph.
3. average degree of a node (weighted and unweighted), which is also among the metrics [Elson et al., 2010] use as it shows how many connections a node (i.e. a single character) has on average in this network.

For further information on metrics we suggest fundamental work by [Wasserman, 1994].

---

[3] Supposedly influenced by Proudhon's *La Guerre et la Paix*, see [Eikhenbaum, 2009 (1931), pp. 498–513]

4. Networks Extraction

In literary networks nodes usually represent characters[4], while edges (and their weights) define some sort of connection or interaction between them. To build a network, one must first formalize this connection somehow. Below we list some of the most common formalizations:

1. Character co-occurrence at certain length. We assume there is an edge between two character nodes if they appear together within the same sentence or paragraph or chapter or simply a text window of a given length. This is the most primitive and abstract formalization, which is nevertheless widely used due to its simplicity. The number of cooccurrences usually becomes the weight of an edge between the characters.
2. Kinship, friendship or any other relations. Explicit mentions of relations in the text are usually quite sparse, and often it makes more sense to build such networks manually. The biggest drawback is that there are usually no weights on the edges, as the relations are mostly binary (relative or not, parent or not, spouse or not). Relation networks usually turn out relatively small and fine-grained, thus limiting the applicability of network measurements.
3. Conversational networks. The two characters are linked each time they engage in a conversation with each other, and the number or length of such conversations becomes the weight of an edge. A more sophisticated subtype of this formalization extends beyond dialogue and accounts for other sorts of social events and interactions as well (one character seeing another, characters engaging in a conflict etc.).

In our work we did not attempt to extract a complex conversational (or social events) network automatically, as this task requires very complicated processing of dialogues and identifying speakers and addressees, who are usually implicit rather than explicit in fiction (for example, [Hee et al., 2013] report that only about 25% of all speech utterances in Jane Austen feature an explicitly named speaker, while 15% have anaphoric reference to the speaker and the remaining 60% are just direct speech with no speaker mentioned at all). In addition to that, in Russian fiction speech instances are often formally indistinguishable from narrative text, as there are no quotations which could serve as formal boundaries. Given all that, we decided to markup interactions between characters in several dozen chapters of Tolstoy's novel by hand (we only marked obvious interactions), and then used these handcrafted conversational networks to evaluate character graphs that we extracted automatically using much more simplistic formalizations of interaction.

Our first set of automatically extracted networks was built on simple co-occurrence of characters in the same sentence. For the second set of networks we tried our own approach based on syntactic structures. The two characters were linked by an edge if they were both syntactic arguments under the same predicate or appeared as two conjuncts (we'll further refer to them as 'syntactic siblings'). It was

[4] Though sometimes locations are also added as separate nodes, see for example [Lee, Yeung, 2012]

our hope that this way we can filter out many 'trashy' connections inevitably made by plain co-occurrence, while still capturing many actual interactions and connections between characters, such as those expressed in examples below:

1) Обедало *человек двадцать, в том числе* ***Долохов*** *и* ***Денисов.***
2) ***он*** *[Николай] вызвал* ***Наташу*** *и спросил, что такое*
3) *Il faut que vous sachiez que c'est une femme, — сказал* ***Андрей Пьеру****.*
4) *Это были* ***Наташа*** *с* ***Соней*** *и* ***Петей****, которые пришли наведаться, не встал ли.*
5) *— Голубчик, Денисов! — взвизгнула* ***Наташа****, не помнившая себя от восторга, подскочила к нему, обняла и поцеловала* ***его****.*

For our experiments we created individual networks for each of the 361 chapter of the novel, as well as bigger and denser aggregated networks for the entire parts (a part in *War and peace* may contain from 12 to 39 chapters). Here is the example of the co-occurrence network for the second part of the first volume of the novel (node sizes proportional to their weighted degrees):

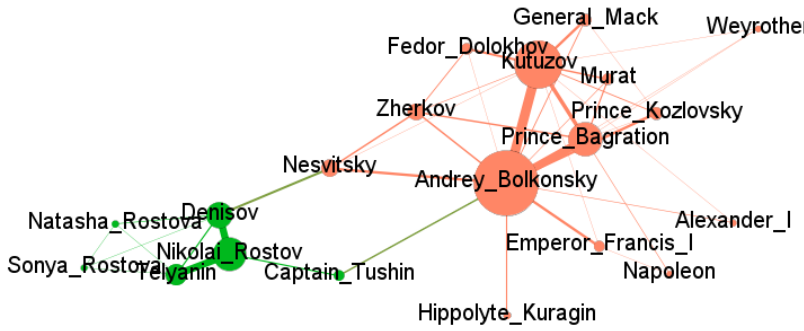

Figure 1. Co-occurrence network for the second part of the first volume of *War and peace*

Of course, before one can extract any kind of a network, character mentions themselves need to be identified throughout the text. This is a challenging task on its own, as it requires named entity recognition (NER), pronominal anaphora resolution and sophisticated nominal coreference resolution (CR). For this task we used a custom extraction model within ABBYY InfoExtractor framework [Stepanova et al., 2016]. This particular model was designed specifically for the task and had a list of *War and peace* character names and aliases. It is important to note that providing the extraction tool with character aliases is also a common practice in digital literary studies of this sort (see, for example, [He et al., 2013]), because so far, no universal NER or CR solution or tool is capable of extracting and linking characters from a random novel with tolerable quality without prior adjustment. This, of course, raises the question of scalability, especially since earlier we claimed that network analysis can be part of a large-scale 'distant reading' approach. We made an attempt to address this issue in the Conclusion and discussion section of this paper.

5. Networks evaluation

5.1. Qualitative evaluation

Before we attempted any quantitative evaluation of the networks, we chose to visualize a number of them to see if they at least 'make sense' at the first sight to someone familiar with *War and peace*. Figures 2 and 3 demonstrate two automatically extracted networks for the entire first part (first 25 chapters) of the first volume of the novel. Here the size of a node is proportional to the weighted degree of that node, and the thickness of an edge reflects its weight, i.e. frequency of co-occurrences in the same sentence or under the same predicate respectively.

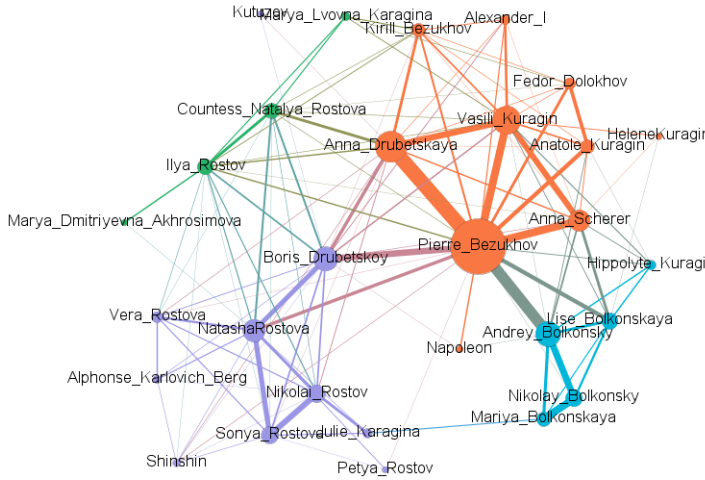

Figure 2. Co-occurrence network, first part of the first volume of *War and peace*

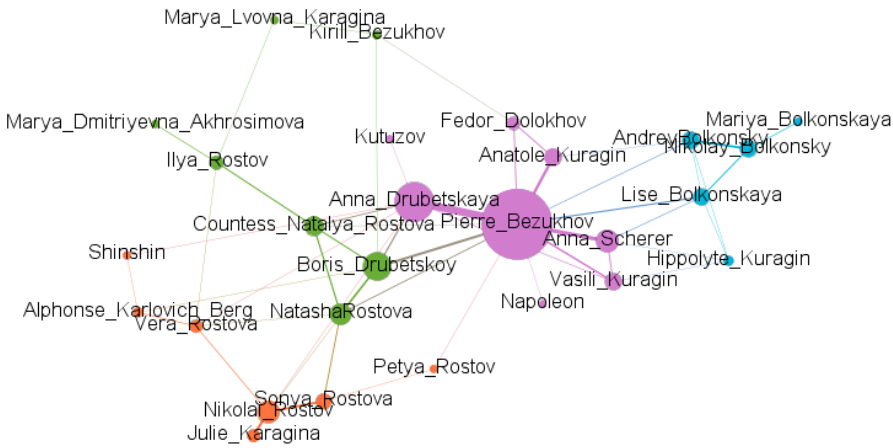

Figure 3. 'Syntactic siblings' network, first part of the first volume of *War and peace*

To a naked eye, both networks look quite similar and seem a pretty adequate reflection of the character system in the first part of the novel. One can easily see that it is centered around Pierre, who appears first at the Anna Sherer's soiree, and then becomes the center of intrigues of Vasili Kuragin and Anna Drubetskaya fighting over the legacy of Pierre's father, count Kirill Bezukhov (and after Vasili Kuragin loses this battle, he strikes back by cajoling Pierre into marrying his daughter Helene). The Rostov family is perfectly visible and, along with their Moscow nobility circle is visibly detached from the St.-Petersburg beau monde which makes up Sherer's soirees. The coloring of the pictures actually reflects automatic modularity clustering made with help of Louvain algorithm [Blondel et al., 2010], and the meaningfulness of the clusters (produced without any adjustments of the default resolution parameters) could also be a sign that the networks reflect certain information about the system of the characters. On Figure 2 one can see four automatically identified clusters:

1. the orange one is mostly St.-Petersburg *beau monde*, which at this point incorporates Pierre, once he turns from a bastard to the new count Bezukhov;
2. the purple one is mainly children and adolescents, the younger generation of Rostov family and Boris (who is, of course, a part of the "Rostov world" at this point in the plot, although already visibly leaning towards the *beau monde* cluster)
3. the turquoise one is the Bolkonsky family, and also Hyppolyte due to his repeated attempts to flirt with Lise. Doing modularity clustering with higher resolution (lesser number of clusters) would connect Andrey, Lise and Hyppolyte to the *beau monde* as well (see Fig.4), while leaving Nikolay Bolkonsky and Mariya Bolkonskaya in their own Bald Hills (Лысые горы) cluster.
4. the green one is the older generation of the Rostov family and their Moscow acquaintances. If we go for lesser clusters (Fig. 4), this one merges with the younger Rostov group.

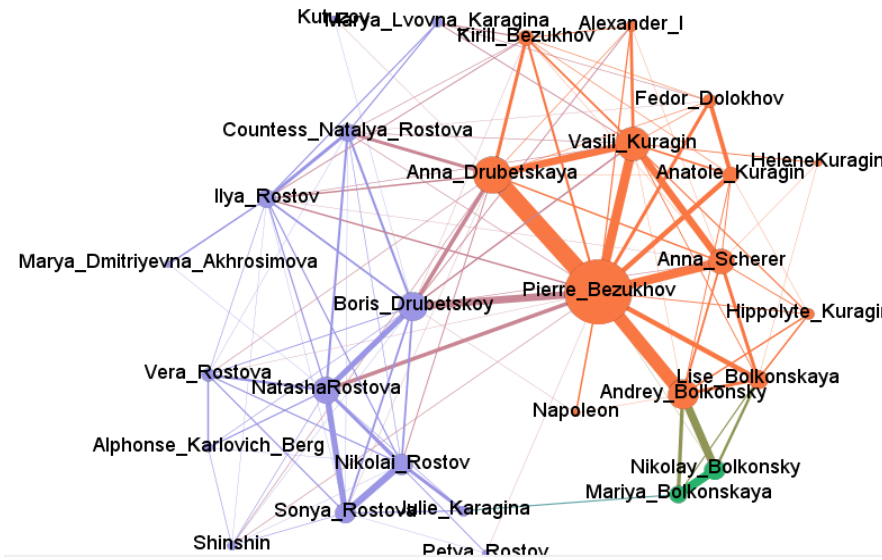

Figure 4. Co-occurrence network, first part of the first volume of *War and peace*

Unlike the “peaceful” first part of the first volume, its second part mainly takes place in the military settings, as the reader follows the experiences of Andrey Bolkonsky at the army headquarters and of Nikolay Rostov in Denisov’s squadron. The networks here (Fig. 5, Fig. 6) are visibly different from Fig. 1–2, not only in their sets of characters, but also in size, density and structure:

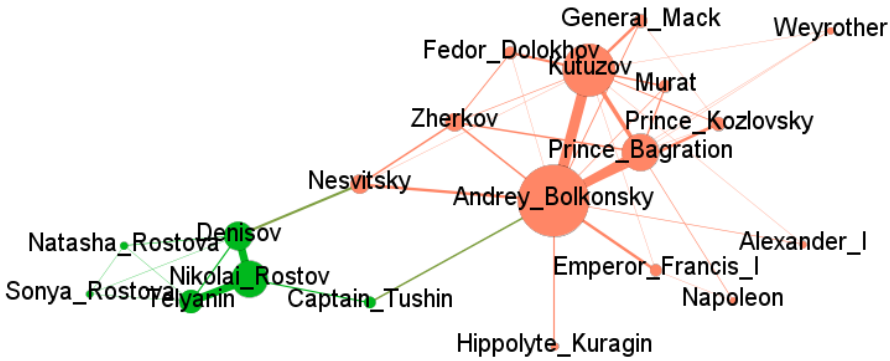

Figure 5. Co-occurrence network, second part of the first volume of *War and peace*

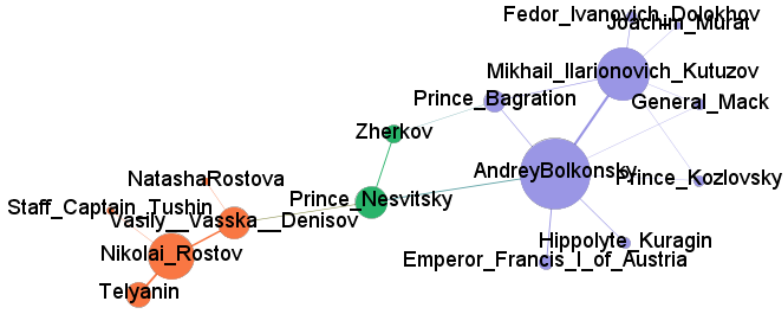

Figure 6. ‘Syntactic siblings’ network, second part of the first volume of *War and peace*

In section 4 of this paper we will try and measure these differences and find out if it occurs on a regular basis between ‘wartime’ and ‘peaceful’ parts of the novel.

The third part of the first volume is essentially a mixture of ‘war’ and ‘peace’ settings (see table 2 in the next section). Pierre spends time with the Kuragin family in St. Petersburg and eventually gets maneuvered into a marriage with Helene, prince Vasily and Anatole pay an unsuccessful visit to Bolkonsky family in the Bald Hills, while Andrey and Nikolay take part in the War of the Third Coalition and both fight in the Austerlitz battle. This heterogeneity and easily distinguishable spatial clusters of part 3 are clearly visible once we plot the graph (again, without any manual adjustments) — see Fig. 7, Fig. 8.

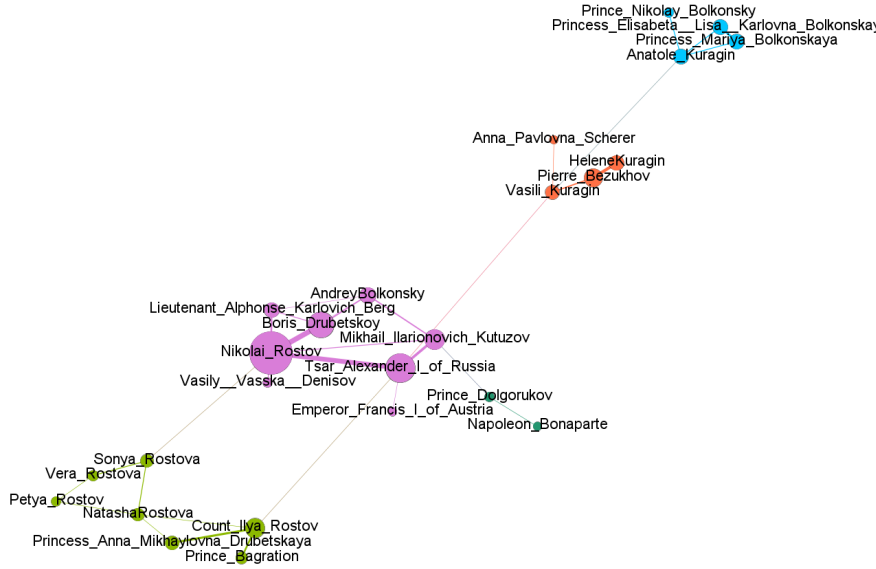

Figure 7. 'Syntactic siblings' network, third part of the first volume of *War and peace*

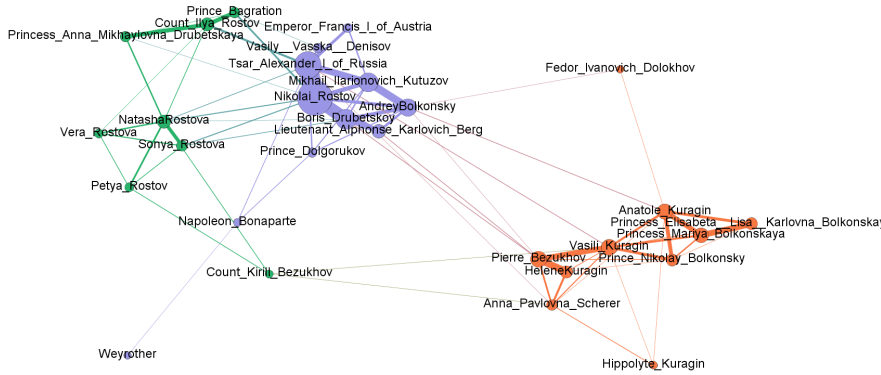

Figure 8. Co-occurrence network, third part of the first volume of *War and peace*

5.2. Quantitative evaluation

We cannot claim that our automatically extracted networks are 'meaningful' and accurately reflect the interactions of characters just because they look like it at the first sight. Therefore we also evaluate them against manually constructed interaction networks for several dozen chapters of *War and peace* (handcrafted networks available at https://github.com/DanilSko/tolstoy/tree/master/Networks/WaP_interactions).

Using Pearson correlation coefficient, we checked which network has more correlation in structural properties to the manually created one across all the chapters. As the results in Table 1 suggest, our 'Syntactic siblings' network outperforms the standard co-occurrence one (the latter in fact has negative correlation in some cases).

Table 1. Correlation of network parameters

Parameter	Correlation with co-occurrence network	Correlation with 'syntactic siblings' network
Density	−0.126	0.840
Diameter	−0.456	0.219
Average degree	0.748	0.923

6. War or peace: testing the hypothesis

As the results of our evaluation suggest, the 'syntactic siblings' network is a much closer approximation of character interaction in a novel than the standard co-occurrence network. Now that we have chosen the method of network extraction, we can finally use it to test our hypothesis. As mentioned above, the idea was to look for correlations between certain parameters of the network and see if they correlate to the kind of settings (army/battlefield or peaceful environments, such as family or high society). We manually classified all 15 parts (excluding the epilogue, which is largely a philosophical essay) of the novel into 'war' (0), 'peace' (1) and 'a mixture of both' (0.5). The results of this classification are shown in the Table 2:

Table 2. Manual classification of 'wartime' and 'peaceful' parts of Tolstoy's *War and peace*

Volume	1	1	1	2	2	2	2	2	3	3	3	4	4	4	4
Part	1	2	3	1	2	3	4	5	1	2	3	1	2	3	4
Peace/ War	1	0	0.5	1	1	1	1	1	0	0	0	0.5	0	0	0

Table 3 shows the parameters of the 'syntactic siblings' network for each part.

Table 3. Parameters of the 'syntactic siblings' network for each part of the novel

Volume	1	1	1	2	2	2	2	2	3	3	3	4	4	4	4
Part	1	2	3	1	2	3	4	5	1	2	3	1	2	3	4
Peace/ War	1	0	0,5	1	1	1	1	1	0	0	0	0,5	0	0	0
Density	0.15	0.16	0.11	0.31	0.25	0.24	0.36	0.21	0.17	0.13	0.13	0.13	0.14	0.18	0.18
Average degree	3.85	2.38	2.64	4.00	2.55	2.67	2.50	3.29	2.00	2.57	2.32	2.00	1.50	1.60	1.60
Average weighted degree	11.41	5.63	7.44	12.86	7.82	6.50	13.00	10.35	5.38	4.76	4.42	3.88	1.67	9.40	5.00

Now we can calculate correlations between the parameters of the 'syntactic siblings' networks extracted from for each part of the novel and the corresponding 'war or peace' value. Table 3 shows the resulting Pearson correlation coefficients.

Table 4. Pearson correlation coefficients between "syntactic siblings" network parameters (first column) and numeric 'war or peace' value we assigned to each part. Positive coefficient means the parameter is statistically higher in 'peaceful' parts

Parameter	Correlation with 'war or peace' value
Density	0.650
Diameter	−0.533
Average degree	0.730
Average weighted degree	0.714

As we can see from Table 4, all parameters have statistically significant correlation with our 'war or peace' target value. Density of the network and its average weighted and unweighted degrees have strong positive correlation with 'peace', quite as our hypothesis predicted. Diameter has a moderate negative correlation with the value, which means that networks with bigger diameter are more likely to be in 'war'-labeled parts.

7. Beyond 'War/peace' antithesis: more prospects for network-based research

The networks we present in this paper were originally created to calculate certain metrics and try to get a quantitative ground for a specific hypothesis. However, they can be used as a novel data-driven model for other research concerning individual characters and their relations to each other in *War and peace.* Some prospects for such research are already foreseeable from the data and visualizations we already have. They are:

- Shifts of the point of view (POV) from one character to another. This is an especially important dimension in research on *War and peace*, as its Tolstoy's trademark technique to show the unfolding events through the eyes and minds of different, constantly shifting characters [Uspensky, 1983], [Bocharov, 1971]. Nodes in the graphs have different sorts of degree and centrality measures which can be used to study he POV changes. Even if we compare our sample graphs for the three parts of the first volume of the (figures 2–8, node sizes proportional to weighted degree), we can see that the central position — and supposedly reader's main viewpoint — is taken first by Pierre, then by Andrey and finally by earlier unimportant Nikolay. Unweighted degree and betweenness centrality of nodes show similar results. Such degree and centrality changes align well with the fact that *War and peace*, when read for the first time by the contemporaries, was often initially perceived as "a novel without main heroes" [Morson, 1987, p.57]. In our networks Natasha becomes central no earlier than the second volume.

- Character groupings. Family unions (the Rostovs, the Bolkonskys, the Kuragins) with all the relations, contrasts and conflicts between them play an extremely important role in *War and peace* [Bocharov, 1971][5]. Relevant to this is the spatial opposition of Moscow vs St. Petersburg circles in 'peaceful' parts and army vs high command in 'war' parts. Without any manual adjustments, our graphs obviously cluster into these groupings (see figures 2–8 again).

8. Conclusion and discussion

We tested two rather simplistic methods for relatively 'low-cost' automatic network extraction from fictional texts and found that the approach based on syntactic structures yields results much closer to manually annotated character interaction networks.

We then used this approach to extract networks from each part of Tolstoy's *War and peace* and test our literary hypothesis. The hypothesis was that 'wartime' parts contain less intensive interaction, which can be approximated by lesser graph densities and average node degrees, as well as bigger diameters. Although all our measurements support the hypothesis, we suggest further, more substantial research before any firm arguments can be made in relation to the composition of the book and authorial techniques behind it.

Apart from this attempt to quantify the differences between 'war' and 'peace' in Tolstoy's novel, our research has revealed other potential applications of network analysis. Namely, we showed that the networks we created could provide insight on POV changes in the narrative and on character groupings and relations.

Another thing that calls for further investigation is the scalability of the approach. One may point out that if a quantitative method is being applied to just one text, however big it is, such method cannot yet be considered successful. But while we admit certain amount of manual work[6], there are two arguments for the general applicability of the approach:

1. There are no fundamental barriers for automating the whole procedure via building and adjusting NER and coreference resolution tools. The fact that state-of-the-art NER/IE/CR applications do not allow seamless transition to XIX century fiction does not imply these texts cannot be handled on a large scale in the nearest future.
2. As more and more semantically and structurally marked digital editions emerge, digital literary scholars eventually become spared from the necessity to process texts with sophisticated NLP machinery. One example is the TEI-encoded corpus of German dramatic texts used by [Trilcke et al., 2015a] for their large-scale (500 texts) network analysis. The XML markup with all the speakers tagged and identified allows easy and reproducible network creation on the go. Currently similar efforts are being made to prepare a TEI edition of Leo Tolstoy's complete works [Skorinkin, 2017].

---

[5] "In *War and peace* family unions, the 'breed' of the character matter a lot. In fact, the Bolkonskys and the Rostovs are more than families — they are separate modes of life" [Bocharov, 1971]

[6] Mainly the adjustment of character list initially obtained from Wikipedia

Acknowledgements

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017 (grant № 17-05-0054) and by the Russian Academic Excellence Project «5-100».

Thy research was partially supported by grant 15-06-99523 from the Russian Foundation for Basic Research.

References

1. *Agarwal A., Kotalwar A., Rambow O.* (2013), Automatic Extraction of Social Networks from Literary Text: A Case Study on Alice in Wonderland, Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan.
2. *Agarwal A., Corvalan A., Jensen J., Rambow O.* (2012), Social network analysis of Alice in Wonderland. Proceedings of the NAACL HLT 2012 Workshop on Computational Linguistics for Literature, pages 88–96, Montreal, Canada.
3. *Alberich, R., Miro-Julia, J., Rossello, F.* (2002), Marvel universe looks almost like a real social network. Preprint arXiv:cond-mat/0202174.
4. *Bocharov S.* (1971), L. N. Tolstoy's War and Peace ["Voina i mir" L. N. Tolstogo], Three masterpieces of Russian classical literature [Tri shedevra russkoi klassiki], Moscow, pp. 7–103.
5. *Bodrova, A., Bocharov, V.,* (2014), Relationship Extraction from Literary Fiction. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], Bekasovo, available at: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/BodrovaAABocharovVV.pdf
6. *Blondel V. D., Guillaume J., Lambiotte R., Lefebvre E.,* (2008), Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment (10), p. 1000
7. *Eikhenbaum, B.* (2009), Works on Leo Tolstoy. Saint-Petersburg. SPBSU Faculty of Philology and Arts.
8. *Elson, D. K., Dames, N. and McKeown, K.* (2010), Extracting Social Networks from Literary Fiction, Proceedings of ACL 2010, Uppsala, Sweden.
9. *Jockers M.* (2013), Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities). University of Illinois Press; 1st Edition.
10. *Lee J., Yeung C. Y.* (2012), Extracting Networks of People and Places from Literary Texts. Proceedings of 26th Pacific Asia Conference on Language, Information, and Computation (PACLIC). pp. 209–218
11. *Moretti F.* (2013), Distant Reading. Verso, London
12. *Moretti F.* (2007), Graphs, Maps, Trees: Abstract Models for a Literary History. Verso, London
13. *Moretti F.* (2011), Network Theory, Plot Analysis. Stanford Literary Lab Pamphlets, Stanford, CA.

14. *Morson G. S.* (1987), Hidden in Plain View: Narrative and Creative Potentials in 'War and Peace'. Stanford University Press, Stanford, CA.
15. *Park, Gyeong-Mi, Kim, Sung-Hwan* (2013), Structural Analysis on Social Network Constructed from Characters in Literature Texts, Journal of Computers, Issue 8
16. *Stepanova, M., Budnikov, E., Chelombeeva, A., Matavina P., Skorinkin, D.* (2016), Information Extraction Based on Deep Syntactic-Semantic Analysis. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2016"], Moscow, pp. 721–733
17. *Skorinkin D.* (2017), Digital Edition of the Complete Works of Leo Tolstoy. 6th AIUCD Conference Book of Abstracts. Rome. pp. 264–267.
18. *Trilcke P., Fischer F., Kampkaspar D.* (2015a), Digitale Netzwerkanalyse dramatischer Texte, in: DHd2015. Von Daten zu Erkenntnissen 23. bis 27. Graz. Book of Abstracts. Austrian Centre for Digital Humanities, 2015.
19. *Trilcke P., Fischer F., Göbel M., Kampkaspar D.* (2015b), Comedy vs. Tragedy: Network Values by Genre. Network Analysis of Dramatic Texts, available at: https://dlina.github.io/Network-Values-by-Genre/
20. *Uspensky, B.* (1983). A Poetics of Composition: The Structure of the Artistic Text and Typology of a Compositional Form. Oakland. University of California Press.
21. *Wasserman, S., and Faust, K.* (1994), Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.
22. *Weitin T.* (2017), Scalable Reading. Zeitschrift für Literaturwissenschaft und Linguistik, Volume 47, Issue 1, pp 1–6